政策解读大数据分析应用的实践探究

摘　要：最新政策和政策解读信息的获取和研究，对于媒体、企业、科研机构等行业都具有重要价值。通过网络数据抓取和文本智能挖掘技术，实现一手政策发布源和解读文章数据源监控采集，并进行数据加工和智能挖掘分析，实现政策解读应用数据产品，可极大提升信息获取和政策研究工作效率。本文从实践角度，介绍政策解读应用中的关键问题解决方法及应用功能。

关键词：政策解读；大数据分析；文本挖掘；数据产品；增值服务

中图分类号：TP393　　**文献标识码：**A

文章编号：1671-0134（2019）03-022-02　　**DOI：**10.19483/j.cnki.11-4653/n.2019.03.001

文 / 谭辛

引言

每年国家各级政府部门都有大量的政策发布，政策内容涉及到社会发展和民生的方方面面，因此每项新政发布都备受关注。政策发布的目的是什么，如何详细解读政策，政策发布后哪些行业领域受到怎样的影响，在政策以及解读已成为媒体界、企业界、学术界，以及普通百姓等社会各方关注的焦点。如何快速和全面地收集和分析政策及解读数据具有现实的意义。

本文介绍通过自动化的网络信息抓取技术、大数据技术和文本智能挖掘技术，快速高效地汇聚最新政策和政策解读信息，并在此基础上进行政策关键要素信息提取、数据多维度分类、数据关联等加工处理，从而形成一套政策解读的数据产品，为媒体从业者、行业研究人员、企业界提供多场景和多角色的信息服务，提升信息获取和数据研究的工作效率。

1. 政策信息获取难点

各行业对最新政策信息获取具有较高的需求和要求。对于媒体行业，需要第一时间获取到政策发布信息，并针对新政撰写宣传报道或解读性文章，稿件发布的时效性体现并影响着媒体的传播力和影响力；对于企业而言，需要及时研究新政对企业发展、研发投入、市场变化、决策分析等方面带来的影响而制定企业内部策略，很多政策的发布对企业发展方向有着至关重要的影响。但政策信息来源广泛、发布时间不集中、信息有待关联整合等制约了政策信息的获取。

1.1 政策数据来源广泛

政策发布均来自各级政府部门，对于个人查询政策信息的难度在于来源广泛的问题。首先，权威的政策查询源头为政府部门的官方网站、官方新闻客户端、官方微信公众号和官方认证微博，发布源头类型较多；其次，政府部门按级别、按部门类型，数量较为庞大，即便仅关注单一领域的政策，也需要关注多个政府官方信息发布源头；再次，对于政策发布后的政策解读文章，除了政府官方网站的官方解读文章外，政府部门官员、领域内专家学者、研究机构的研究人员、媒体业专家等撰写的解读文章也具有非常高的阅读价值，但是这些来源更为广泛，可能来自比如新闻门户网站、新闻客户端、报纸或纸媒电子报、机构的微信公众号或微博、领域内专家学者个人的博客、公众号或微博等。总之，想要快速浏览到各方发表的政策解读信息存在一定困难。

1.2 发布时间不一

每年发布的政策中，只有很少一部分是在固定时间段内发布，另外的大部分都是根据社会发展需要而实时推出的，无法提前准备政策信息获取工作。

综上所述，在信息过载的当下，如何快速高效获取政策和解读信息，如何精准获取各行业研究人员需要的数据，如何借助人工智能和机器分析能力汇聚分析信息为研究人员服务，成为政策解读应用需要解决的关键问题。

2. 政策解读大数据分析的应用实践

政策解读应用借助大数据和人工智能技术，实现了自动化的信息采集、多维度的自动标引、文本挖掘和关联分析，通过可视化的展示提升了政策及政策解读信息的使用效率。

2.1 自动化监测采集

通过借助成熟的自动化网络信息抓取软件，实现对政策和解读信息发布源头目标网站做实时监控，把最新的网页及时采集到本地，进行内容分析和信息过滤等流程，完成政策解读信息本地存储。

数据采集过程中，应用不仅将网页的非结构化数据转变成半结构化数据，同时自动提取政策名称、发布时间、政策文本内容，以及发文单位名称、发布网站名称、频道名称、发文链接地址等政策相关数据。后续进行的文本挖掘和加工处理，构建了政策元数据数据库，为政策解读应用提供基础数据服务。

采集源头主要面向一手发布数据源，而非经过转载后的二手数据，以保证信息获取的及时性、准确性和可靠性。

2.2 政策和解读信息加工处理

数据采集技术，对最新网络数据实时监控采集，解放个人浏览和搜索时间。文本挖掘技术，提供信息自动

化分类、自动聚类、智能化信息提取、数据关联分析和数据自动标引等一系列数据加工处理，解决政策数据孤岛问题，让政策数据应用更加有效。图 1 为政策及解读数据加工处理流程图。

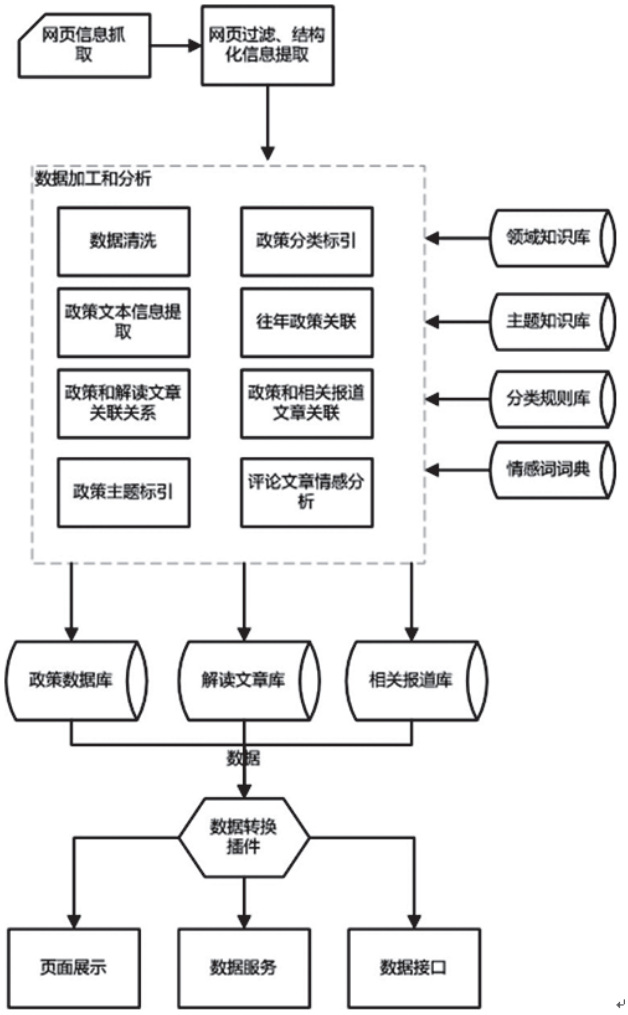

图 1　数据加工处理流程图

2.3 多维度分类标引

采用自动分类和规则分类技术，对政策做多维度分类标引，用以帮助不同身份用户在不同需求场景下更加快速、有针对性地查找到所需类目和对应的政策信息。包括政策所属行业领域、所属地域名称、发布单位名称、所属主题名称、发文形式、所属年份等分类标签。分类类别举例如表 1 所示。

表 1　政策分类名称和分类值举例

分类名称	分类值举例
行业领域	金融、制造、互联网……
地域	北京、山东、湖北……
发布单位	国务院、国家发改委、工信部……
主题	金融监管、个人所得税、电子商务……
发文形式	公告、通知、议案……
年份	2019、2018、2017……

在前端应用功能中，利用这些政策标签，采用细分导航的方式，进行政策列表展示。通过组合式的检索功能对政策和解读文章进行搜索，让用户可以通过自定义关键词的方式获取个性化的检索结果，达到快速、全面了解信息的目的。对政策和解读文章的标题、正文和主题提供全文检索功能。对政策的发布单位名称、发文形式、所属行业领域、所属地域、发布年份等字段，提供筛选功能。多维度分类标签，也让页面筛选更为灵活，为个性化订阅提供基础选项。

2.4 政策文本挖掘

对政策文本做数据挖掘和关键信息提取，是政策索引和检索、信息关联分析、多维度分类标引等数据加工的基础。采用文本自动分词和词性标注等自然语言处理技术，基于规则与统计相结合的方式，将政策文本进行中文分词以及政策信息提取，包括政策主题关键词、相关人物、机构、地区名称等信息的结构化提取，完成政策的关键词和实体标引。

在政策信息展示功能中，通过多维度的智能分析与关联，帮助用户快速地发现该政策中的关键信息以及关联文章。以图表化形式，展示政策主题词、政策主体挖掘结果（相关人物、相关机构、相关地区）、政策解读文章时间发布趋势和数量；以文章标题列表方式，展示相关政策、相关解读文章、相关媒体报道文章。展现结果示意图如图 2 所示。

图 2　政策挖掘结果展示图

快速挖掘多方观点，对多方观点进行对比展示，可以使用户更全面地把握政策内容。利用语义分析技术，把多文章之间的相关度超过一定阈值的文章关联到一起，实现复杂语义关系的深度挖掘，从而完成政策与官方解读文章、媒体解读文章、相关报道文章、往年政策等进

（下转第 38 页）